图书情报工作
第63卷 第15期 2019年8月

中国政府网"一带一路"新闻话题挖掘分析*

秦玥[1] 吴亚平[2] 王继民[1]
[1] 北京大学信息管理系 北京 100871 [2] 北京大学图书馆 北京 100871

摘要:［目的/意义］探究中国政府网"一带一路"相关新闻的话题内容及热度变化,呈现"一带一路"倡议主题及动态,明确不同时期的倡议重点,为相关研究提供参考。［方法/过程］构建基于LDA模型的新闻话题内容的基本框架,限定2015－2017年"一带一路"相关新闻数据,利用LDA模型进行话题抽取,根据文档与话题的概率分布计算,分析各主题在不同时间段的热度演化。［结果/结论］抽取得出30个细分话题,归纳为政策沟通、设施联通、贸易畅通、资金融通、民心相通、"一带一路"对我国经济的影响和政府工作7大类。其中,政策沟通类在全时间段上热度最高,贸易畅通和"一带一路"对我国经济的影响两类话题紧随其后。"进出口"等细分话题的热度不断上升,"改革与转型"等细分话题的热度则有下降,体现了官方媒体新闻内容及其关注度随时间而变化的特点。

关键词:"一带一路" LDA模型 话题抽取 热度演化
分类号: G250
DOI: 10.13266/j.issn.0252－3116.2019.15.012

1 引言

2015年3月,国家发展改革委、外交部、商务部联合发布了《推动共建丝绸之路经济带和21世纪海上丝绸之路的愿景与行动》,"一带一路"倡议正式进入实施阶段[1]。"一带一路"倡议以政策沟通、设施联通、贸易畅通、资金融通、民心相通为重点合作内容,自提出以来已经引起了国内外高度关注与热烈反响,我国官方媒体对该倡议进行了热烈报道。对"一带一路"相关新闻内容进行话题分析,有助于公众了解"一带一路"倡议的动态及不同时期的倡议重点,进一步丰富"一带一路"相关研究。笔者选取具有权威性的"中华人民共和国中央人民政府门户网站"(以下简称"中国政府网")的"一带一路"相关新闻数据,利用LDA模型进行话题抽取,得出30个细分话题,将其归纳为政策沟通、设施联通、贸易畅通、资金融通、民心相通、"一带一路"对我国经济的影响和政府工作七大类别,计算、分析了七大类别和细分话题在不同时间段的热度演化,按热度变化情况将其分为上升类、下降类和波动类三类话题,呈现了官方媒体对"一带一路"不同新闻内容关注度的变化情况。

2 相关研究

2.1 "一带一路"新闻研究

目前,围绕"一带一路"新闻的研究以国内为主,集中在新闻报道框架研究和新闻内容的量化分析两方面。新闻框架是指新闻媒体在对新闻事实进行选择性处理时所依据的特定原则。新闻报道框架研究通常是指对新闻篇幅、新闻来源、新闻选题等进行分析,以总结出新闻报道的框架特色。如:姚玉娇选取《人民日报》"一带一路"相关报道,从新闻材料的选取与建构、报道内容、报道主题等方面探讨了"一带一路"新闻报道的生产过程、信息框选和事实建构特点,认为《人民日报》形成了以正面引导为主、重视政策宣传和成就展示的新闻框架[2];曾润喜等选取人民网、新华网等18家主流媒体的118篇"一带一路"相关新闻报道,从报道视角、报道内容、报道体裁和语言风格几方面进行分析,发现当前新闻内容存在政宣口气浓重、关注点雷同、单一强调中国作用等问题,阻碍了社会对"一带一路"的正确认识[3]。

* 本文系国家社会科学基金项目"'一带一路'沿线国家互联互通水平综合评价研究"(项目编号:16BTQ057)研究成果之一。
作者简介:秦玥(ORCID:0000－0002－1948－5904),硕士研究生;吴亚平(ORCID: 0000－0002－4242－2434),馆员;王继民(ORCID:0000－0002－3573－7788),教授,博士生导师,通讯作者,E-mail:wjm@pku.edu.cn。
收稿日期:2018－12－11 修回日期:2019－03－27 本文起止页码:103－110 本文责任编辑:易飞

针对新闻内容的量化分析研究通常更关注新闻内容本身,试图探究新闻文章的主题内容。如:汪海藻选取了114篇《中国日报》和466篇《中国日报·美国版》的"一带一路"相关新闻,通过词频统计方法提取关键词并划分类别,发现《中国日报》报道内容侧重于经济类别,《中国日报·美国版》侧重于综合类报道[4];田作宇选取了594篇"一带一路"相关的印度英语新闻,将语料库话语分析与情感词典相结合,运用词表分析、历时主题词分析、词簇分析等方法,探究新闻对"一带一路"倡议的解读和评价,发现印度新闻报道的关注内容主要集中在领导人互访、中国与印度邻国的相关合作、亚投行成立等问题上,媒体态度包含怀疑和揣测等多种复杂情绪[5]。

总体来看,目前"一带一路"新闻相关研究以新闻报道框架研究为主,针对新闻内容的量化分析研究数量较少,并且量化分析多使用词频统计、词簇分析等方法,较难对新闻话题进行深入挖掘。

2.2 新闻话题抽取及演化分析研究

话题检测与跟踪技术(topic detection and tracking,简称TDT)是对新闻媒体信息流进行新话题自动识别和对已知话题进行持续跟踪的技术,已成为信息爆炸时代信息处理领域的热点技术之一。对新闻话题进行话题抽取和演化分析是TDT的应用之一,常见的新闻数据建模方法包括基于向量空间模型的方法、基于语言模型的方法和基于概率主题模型的方法等。

(1)向量空间模型(vector space model,简称VSM)由G. Salton等于20世纪70年代提出[6]。该模型将文档表示为向量,将对文本内容的处理简化为向量空间中的向量运算。J. Allan等以广播新闻报道为数据源,将新闻报道表示为特征向量,利用VSM找出若干新闻话题对应的特征向量,判断新报道的文章属于已知话题还是新话题,实现话题的检测和追踪[7]。林南根据新闻报道的结构和时间特征,提出了适用于话题检测的TD-VSM模型,使用信息熵和新闻报道的结构特征改进TF-IDF权重计算,结合新闻报道的时间特征改进余弦相似度计算,实现对新闻话题的识别[8]。

(2)语言模型(language model,简称LM)由M. Spitters于2002年首次提出[9]。该模型根据语言客观事实进行语言抽象数学建模,包括N-gram模型、决策树模型等。V. Lavrenko等利用特殊的一元语言模型,即相关性模型,对已有的话题相关新闻文档进行动态信息扩充,提高了话题模型的信息全面性[10]。C. Zhai等研究了语言模型平滑问题及其对检索性能的影响,发现检索性能不仅对平滑参数敏感,而且灵敏度模式受查询类型的影响,具有性能优势[11]。

(3)概率主题模型(probabilistic topic model,简称PTM)的理论思想起源于T. Hofmann在潜在语义分析(latent semantic analysis,简称LSA)基础上提出的概率隐性语义分析模型(probabilistic latent semantic analysis,简称pLSA),该模型认为每篇文档由话题的多项式分布随机生成,不同话题又会产生不同的词[12]。D. M. Blei等在2003年提出了LDA模型(latent dirichlet allocation),它是一个三层贝叶斯生成概率模型,该模型将文档集合模拟为潜在话题的有限混合,潜在的话题集合又由若干个特征词汇构成[13]。之后出现了许多改进与扩展后的概率主题模型,如考虑时间因素的动态主题模型[14]等。L. Alsumait等提出改进的在线主题模型OLDA,在线自动识别新出现的新闻文档的新增主题,根据新数据流推断的信息增量式地更新主题模型,及时掌握各个主题随时间的变化情况[15]。楚克明等以两会新闻为例,提出一种挖掘新闻话题随时间变化的方法,先利用LDA模型对不同时间段的文档集合进行话题抽取,再计算相邻时间段中的任意两个话题的分布距离,以发现话题之间的内容关联,得出新闻话题的内容演化[16]。

除上述三类模型外,词汇链模型、图模型等方法也有所出现,不断丰富着新闻话题抽取与演化分析的相关研究。总体来讲,向量空间模型虽然应用广泛,但由于没有考虑文字之间的语义关联,仍存在一定缺陷;语言模型在突发性的新闻话题上欠缺一定的准确性,尚未成为主流;概率主题模型具有较好的泛化性,并可以通过扩展模型,使其能在处理短文本等方面也取得不错的效果,正在被广泛应用。目前对于"一带一路"新闻的研究主要集中于新闻框架研究,缺乏对新闻内容本身的深入研究。作为概率主题模型的一种,LDA模型具有强大的话题识别能力,已被应用于话题发现、文本分类、文本聚类、情感分析等多个领域,有着较好的效果。因此,本文选取LDA模型分析"一带一路"相关新闻,以探究新闻话题的构成及其热度演化情况。

3 话题抽取与演化分析框架

3.1 话题抽取框架

LDA模型是一个三层贝叶斯生成概率模型,包含词、主题和文档三层结构,将文档模拟为潜在话题的有限混合。在LDA模型的三层结构中,首先假设词是由话题的概率分布混合而成,再假设文档由潜在话题的

概率分布混合而成。对于每篇文档,先从 Dirichlet 分布中抽样产生该文档包含的话题比例,再结合话题和词的概率分布生成文档中的每一个词[16]。如下步骤详细描述了 LDA 模型中一篇文档的生成过程,其中使用的符号及其含义见表 1。

(1)对文档集合中的文档 d,根据 $\theta_d \sim Dirichlet(\alpha)$ 生成该文档上的话题分布;

(2)文档 d 中第 i 个词 w_{di} 的生成:

- 生成一个话题 $z_k \sim Multinomial(\theta_d)$;
- 根据 $\varphi_k \sim Dirichlet(\beta)$ 生成话题在词表上的分布;
- 生成使 $p(w_{di} | \varphi_k)$ 最大的一个词。

表 1　LDA 模型符号含义说明

符号	含义
d	一篇文档
w_{di}	文档 d 的第 i 个词
z_k	话题 k
θ_d	文档 d 的话题的多项式分布
φ_k	话题 k 在词表上的多项式分布
α	文档－话题分布的先验参数
β	话题－词分布的先验参数

LDA 模型引入了 α 和 β 来完成文档的生成过程,并通过 Gibbs Sampling、期望扩散等方法来对 θ 和 φ 两个参数进行近似推理,得到文档的话题。该方法的关键在于如何求解当前词语采样的概率,其求解得到的 θ 和 φ 的后验估计值[17]表达式如下:

$$\hat{\theta}_{dj} = \frac{C_{dj}^{DK} + \alpha}{\sum_{j=1}^{K} C_{dj}^{DK} + K\alpha} \qquad \hat{\varphi}_{ij} = \frac{C_{ij}^{VK} + \beta}{\sum_{i=1}^{V} C_{ij}^{CK} + K\beta}$$

其中 K 表示话题数目,C_{dj}^{DK} 表示文档 d 中指派给第 j 个话题的词数目,$\sum_{j=1}^{K} C_{dj}^{DK}$ 表示文档 d 中所有被分配了话题的词数目,C_{ij}^{VK} 表示第 i 个词指派给第 j 个话题的次数,$\sum_{i=1}^{V} C_{ij}^{VK}$ 表示指派给第 j 个话题的所有词数目。

本文沿用 LDA 模型中对话题的定义,即话题是一组语义相关的词语及这些词语在该话题上的分布概率值,可表示为:

$Z = \{(w_1, p(w_1 | z)), (w_2, p(w_2 | z)), \cdots, (w_V, p(w_V | z))\}$

其中,Z 表示话题,w_i 表示第 i 个词语,$p(w_i | z)$ 表示话题 Z 下出现第 i 个词语的概率值,V 表示词表的大小[18]。

利用 LDA 模型对新闻文档进行话题抽取时,抽取的话题数目的设定非常关键。话题一致性(topic coherence)[19]衡量了某一话题下高频出现概率词语之间的语义相似程度,可用于 LSA、LDA 等模型的评估。对于模型中的话题,如果该话题下高频出现概率词语间的语义相似程度较高,则认为该话题的一致性较高,模型效果较好。将话题数目设置为等距的多个值(N_1, N_2, $\cdots$ N_n),计算每个话题数目下模型的话题一致性程度,话题一致性的最高值对应最优话题数目,本文选取此指标作为话题数目选取的依据。

话题抽取流程如图 1 所示,即采集得到中国政府网中“一带一路”相关新闻文档集合后,对数据集进行数据清洗、分词等预处理过程,根据话题一致性指标选取最佳话题数目,再利用 LDA 模型抽取话题,并对话题内容进行类别划分,实现新闻文档的内容挖掘。



图 1　话题抽取框架

3.2　话题热度演化分析框架

话题热度一般通过话题与文档之间的关联度表示,同一话题可能以不同的重要程度出现在各个文档中,一个话题被越多文章提及,则热度越高。通过计算某一话题在不同时间段内的热度,可反映话题热度随时间变化的趋势,实现话题热度的演化分析。话题的热度根据文档－话题的分布计算,即计算得出某个话题在所有文档中出现概率的平均值,如话题 z_k 在某一时间段中的热度可以表示为:

$$\delta_k = \frac{\sum_{d \in D} \theta_{dk}}{|D|}$$

其中,D 表示某一时间段中的文档集合,|D|表示文档集合 D 中的文档数量,d 表示 D 中的一篇文档,θ_{dk} 表示话题 z_k 出现在文档 d 中的概率。

对所有新闻文档集合抽取话题后,首先计算得出各个话题在全时间段上的总体热度排名,之后将文档集合按照其发布时间离散到各个时间窗口中,利用 LDA 模型得到的文档－话题分布矩阵计算各个话题在每个时间窗口内的热度,得到话题热度随时间的变化情况。根据每个话题的热度走势,将其按照热度变化划分为上升类话题、下降类话题和波动类话题,得到不同话题的热度演化情况。具体话题热度演化分析框架见图 2。

4　实验过程及结果分析

4.1　数据采集

分别以“一带一路”“丝绸之路经济带”“21 世纪

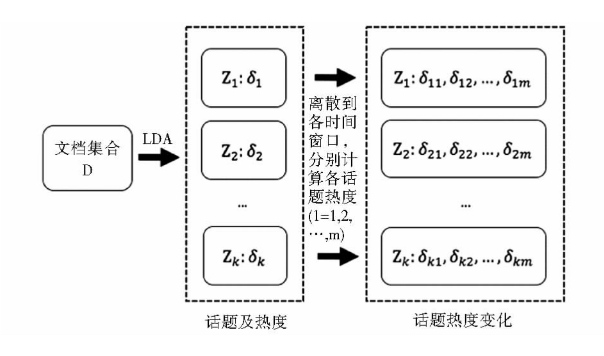


图 2　话题热度演化分析框架

海上丝绸之路”作为检索关键词，将时间限定为 2015－2017 年间，对检索结果去重后得到 8 069 篇新闻文档，年度数量分布如表 2 所示：

表 2　“一带一路”相关新闻篇数统计

年份	2015 年	2016 年	2017 年
新闻篇数	2428	3019	2622

为提高实验准确性，对采集到的初始数据进行预处理，包括：去除新闻文本中无意义的字符；将“一带一路”相关词语添加到用户自定义词典中，防止其被错误划分；利用结巴分词对新闻语料进行分词处理并过滤语料中的停用词以及人名等对话题区分度不高的词语等。

4.2　话题抽取及演化分析结果

4.2.1　话题抽取结果　以话题一致性指标为衡量标准，通过实验的方法选择合适的话题数目，实验结果如图 3 所示：

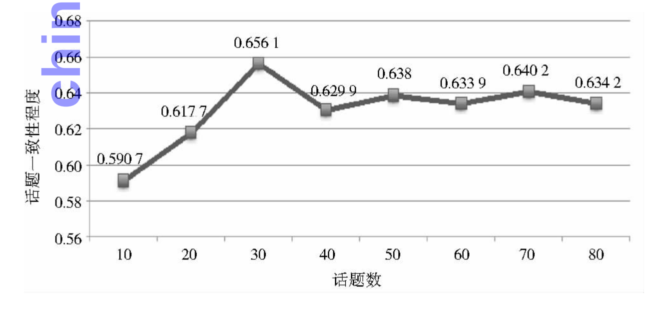


图 3　不同话题数下的话题一致性程度

可以看出，话题数为 30 时，话题中词语的语义相似程度最高，话题一致性最强，因此将话题数目确定为 30 最佳，通过 LDA 模型共抽取出 30 个与“一带一路”相关的新闻话题。笔者结合话题关键词以及话题对应的新闻文档内容，将 30 个话题归纳为七个大类，分别是：政策沟通、设施联通、贸易畅通、资金融通、民心相通、“一带一路”对我国经济的影响和政府工作。具体对应的细分话题、话题词、话题热度及排名如表 3 所示：

可以看出：①政策沟通类别涵盖“一带一路”沿线国家政府层面的合作互通，体现各国政府如何沟通、交流，以达成政治互信与合作共识，共包含领导人会谈等 7 个话题；②设施联通类别涵盖“一带一路”沿线国家的基础设施建设，形成连接沿线各国的基础设施网络，共包含交通建设等 3 个话题；③贸易畅通类别涵盖“一带一路”沿线国家的投资与贸易合作，致力于解决投资与贸易便利化的问题，共包含产业创新等 6 个话题；④资金融通类别涵盖“一带一路”沿线国家的金融合作与监管，包含跨境金融服务话题；⑤民心相通类别涵盖“一带一路”沿线国家对丝绸之路友好合作精神的传承，在文化交流、学术往来等各个领域展开合作交流，共包含科研创新等 6 个话题；⑥“一带一路”对我国经济的影响类别不以国家间的合作互通为核心，而是着眼于我国经济在“一带一路”倡议后的变化与发展，共包含改革与转型等 4 个话题；⑦政府工作类别更关注我国政府在“一带一路”倡议下的工作内容，共包含制度管理等 3 个话题。

总体热度排名前 5 位的细分话题分别是：改革与转型、领导人会谈、产业创新、博鳌亚洲论坛、交通建设，具体含义如下：

（1）改革与转型。该话题阐述了“一带一路”对我国经济转型的影响。“一带一路”倡议具有提升投资贸易便利性、优化贸易结构、提升科技水平等优势，可推进我国经济的自由化、市场化和国际化，并加重科技力量在经济中的作用，促进我国经济结构的调整。

（2）领导人会谈。该话题主要涉及“一带一路”沿线各国领导人的出访、会谈、贺电等新闻，描述各国领导人如何沟通“一带一路”合作并建立起互利共赢的全面战略伙伴关系。领导人之间的接触是国家间沟通的直接桥梁，是“一带一路”沿线国家政策互通的重要环节。

（3）产业创新。该话题提及我国长江经济带、长三角城市群、长吉新区等产业创新园区的建立和发展。这些园区的建立有助于我国发挥产业优势、促进产业创新，从而作为节点地区来推动“一带一路”沿线各国的产业升级与合作，促进经贸发展。

（4）博鳌亚洲论坛。该话题涉及的博鳌亚洲论坛是由 25 个亚洲国家和澳大利亚共同发起的国际会议组织，旨在增进亚洲各国之间、亚洲各国与世界其他地区之间的交流与合作。近几年来，随着“一带一路”倡议的不断发展，该倡议已逐渐成为博鳌亚洲论坛的议题之一，在会议中占据一定的比重。博鳌亚洲论坛已

表 3 "一带一路"相关新闻话题结果展示

类别	话题	抽取出的话题词(前 10 个)	话题热度	热度排名
政策沟通	领导人会谈	交流,深化,会见,务实,互利,友好,人文,签署,战略伙伴,访问	0.063	2
	博鳌亚洲论坛	亚洲,论坛,非洲,亚投行,基础设施,博鳌,未来,关注,中非,媒体	0.061	4
	会议召开	新疆,论坛,博览会,交流,会议,介绍,主任,部长,援疆,召开	0.045	7
	人类命运共同体	治理,理念,开放,共同体,命运,历史,大国,外交,构建,东亚	0.044	8
	世界局势探讨	安全,会议,联合国,上合,成员国,稳定,维护,问题,峰会,对话	0.042	11
	同中东欧国家的合作	中欧,中东欧,俄,欧洲,沿线,俄罗斯,欧亚,欧盟,贸易,签署	0.038	12
	同东盟国家的合作	东盟,海上,马来西亚,世纪,东盟国家,共同体,越南,泰国,互联互通,区域	0.032	15
设施联通	交通建设	铁路,物流,高铁,交通,港口,运输,交通运输,公路,通道,班列	0.049	5
	能源建设	能源,产能,装备,制造,技术,产业,生产,产品,核电,集团	0.026	18
	通信建设	海洋,印尼,海上,世纪,卫星,国家海洋局,福建,测绘,福州,航天	0.014	27
贸易畅通	产业创新	产业,优势,城市,区域,全国,中心,全省,长江,平台,基地	0.062	3
	进出口	出口,外贸,贸易,进出口,对外,进口,商务部,直接,沿线,下降	0.043	9
	自贸试验区	自贸区,开放,贸易,试验区,海关,通关,上海,对外开放,自贸,区域	0.026	19
	税务及审计	政府,审批,部门,税收,政策,地方,试点,管理,行政,资金	0.025	21
	港澳的经贸角色	香港,谈判,协定,经贸合作,贸易,经贸,内地,自贸区,澳门,部长	0.023	22
	产权保护	互联网,信息,知识产权,电子商务,平台,信息化,网络,品牌,数据,质检	0.023	23
资金融通	跨境金融服务	金融,融资,人民币,银行,资本,基金,资金,风险,金融机构,贷款	0.025	20
民心相通	科研创新	科技,技术,标准,研究,人才,标准化,体系,能力,资源,研发	0.032	14
	文化交流	文化,民族,宗教,传统,出版,传播,文明,优秀,海外,精神	0.018	25
	旅游产业	旅游,两岸,海南,台湾,游客,旅游业,大陆,海南省,岛,邮轮	0.015	26
	人才培养	创业,教育,就业,高校,人才,培训,计划,青年,少数民族,学校	0.014	28
	生态建设	生态,保护,林业,森林,文明,气候变化,绿色,治理,防治,面积	0.008	29
	中医药领域的合作交流	医疗,健康,中医药,卫生,医院,体育,甘肃,中医,药品,医药	0.004	30
"一带一路"对我国经济的影响	改革与转型	未来,转型,政策,当前,常态,需求,政府,经济体,面临,巨大	0.067	1
	区域协调发展	规划,区域,协调,目标,重大,民生,生态,经济社会,基础设施,水平	0.045	6
	经济效益	增速,消费,百分点,下降,保持,工业,提高,服务业,以上,结构	0.038	13
	农业供给侧改革	农业,供给,农产品,产能,结构性,侧,粮食,政策,提高,农村	0.020	24
政府工作	制度管理	制度,气象,管理,意见,完善,落实,要求,部门,保障,安全	0.042	10
	中央精神	总书记,委员,党,精神,中央,报告,全国政协,党中央,领导,同志	0.028	16
	政府发展方向	问题,做,解决,比较,政府,情况,包括,环境,应该,过程	0.027	17

经成为亚洲各国政策互通的国际性平台。

(5)交通建设。该话题涉及中欧班列等铁路建设、陆水联运通道口岸建设等内容,体现了"一带一路"沿线国家为提升道路通达水平、实现全面畅通的国际物流运输所做的努力。

4.2.2 话题热度演化分析结果 7 个类别的话题在全时间段上的热度情况见图 4。

其中,政策沟通类别的话题热度最高,热度值为 0.32,几乎占据了总热度的三分之一;贸易畅通类别的话题热度排第二位,热度值为 0.20;"一带一路"对我国经济的影响类别的话题热度排第三位,热度值为0.17。其他 4 类话题的热度值则相对较低,均不超过0.10。

按照新闻文档的发布时间,将文档集合按照季度离散,共包括从 2015 年第一季度至 2017 年第四季度 12 个时间段。分别计算各个话题在不同时间段中的热度,得到话题的热度演化结果。整体来看,7 个类别的话题在不同时间段的热度值结果见图 5。

图 5 显示,政策沟通类别的话题热度波动较大,但始终占据所有类别话题中的热度最高值。贸易畅通和"一带一路"对我国经济的影响类别的话题热度占据第二到三位,二者热度最初较为接近,2016 年第二季度后贸易畅通类别的话题热度更具优势。其他类别的话题热度相对较低,且没有明显的波动。

30 个细分话题在不同时间段的热度值及总体热度走势如表 4 所示。其中,数值背景颜色的深浅表示热度的高低,颜色越深热度越高,颜色越浅热度越低。箭头与横线表示话题的热度走势,指向右上方的箭头表示该话题的热度为上升趋势,指向右下方的箭头表示该话题的热度为下降趋势,横线表示话题热度平稳波动,没有明显的上升或下降趋势。

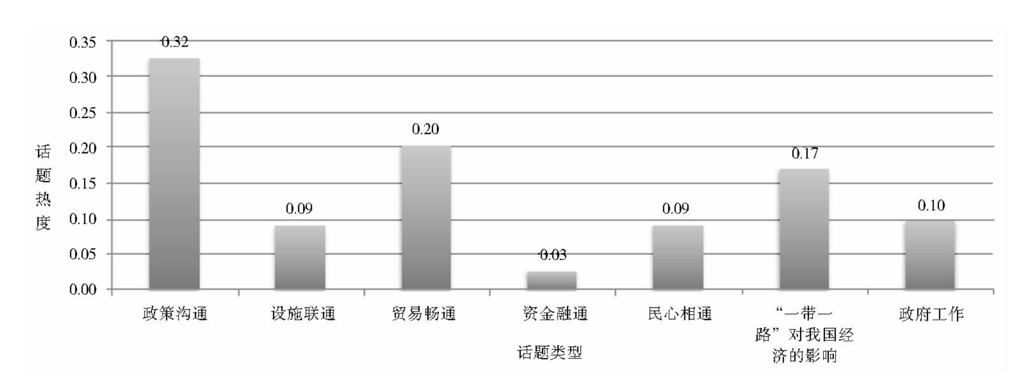


图 4　“一带一路”各类话题热度

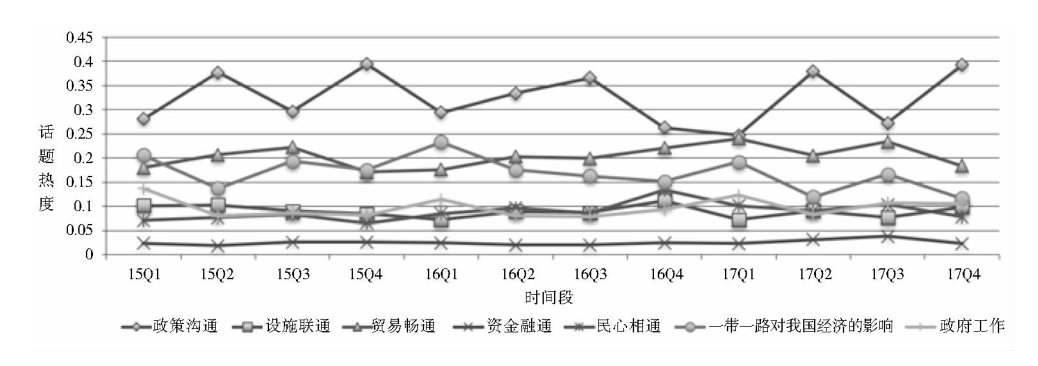


图 5　“一带一路”各类话题热度变化

表 4　“一带一路”细分话题热度变化

类别	话题	15Q1	15Q2	15Q3	15Q4	16Q1	16Q2	16Q3	16Q4	17Q1	17Q2	17Q3	17Q4	趋势
政策沟通	领导人会谈	0.039	0.102	0.050	0.066	0.049	0.078	0.059	0.050	0.050	0.095	0.039	0.080	–
	博鳌亚洲论坛	0.082	0.072	0.051	0.066	0.084	0.054	0.058	0.045	0.066	0.065	0.044	0.048	↘
	会议召开	0.028	0.050	0.057	0.029	0.022	0.051	0.052	0.056	0.028	0.064	0.067	0.036	↗
	人类命运共同体	0.048	0.027	0.034	0.058	0.047	0.027	0.059	0.033	0.039	0.044	0.043	0.068	–
	世界局势探讨	0.030	0.036	0.038	0.066	0.034	0.051	0.056	0.026	0.026	0.047	0.033	0.053	–
	同中东欧国家的合作	0.027	0.059	0.030	0.060	0.028	0.047	0.038	0.028	0.014	0.044	0.021	0.071	–
	同东盟国家的合作	0.027	0.030	0.037	0.049	0.031	0.027	0.043	0.026	0.025	0.021	0.024	0.037	–
设施联通	交通建设	0.059	0.059	0.049	0.050	0.035	0.042	0.051	0.060	0.037	0.052	0.042	0.058	–
	能源建设	0.027	0.027	0.027	0.024	0.029	0.032	0.021	0.032	0.024	0.026	0.025	0.021	–
	通信建设	0.016	0.017	0.015	0.010	0.009	0.016	0.015	0.020	0.012	0.013	0.010	0.019	–
贸易畅通	产业创新	0.067	0.064	0.083	0.048	0.053	0.056	0.060	0.085	0.063	0.050	0.070	0.048	–
	进出口	0.027	0.032	0.045	0.035	0.030	0.064	0.046	0.046	0.052	0.047	0.063	0.036	↗
	自贸试验区	0.030	0.033	0.028	0.027	0.026	0.019	0.025	0.019	0.035	0.028	0.020	0.025	–
	税务及审计	0.030	0.031	0.028	0.017	0.021	0.022	0.024	0.021	0.031	0.026	0.034	0.025	–
	港澳的经贸角色	0.014	0.027	0.019	0.027	0.028	0.018	0.019	0.027	0.030	0.024	0.016	0.025	–
	产权保护	0.012	0.021	0.021	0.016	0.018	0.024	0.024	0.023	0.030	0.030	0.031	0.024	↗
资金融通	跨境金融服务	0.024	0.019	0.026	0.027	0.025	0.020	0.020	0.025	0.023	0.031	0.039	0.023	–
民心相通	科研创新	0.019	0.024	0.027	0.025	0.029	0.037	0.036	0.047	0.034	0.035	0.033	0.032	↗
	文化交流	0.017	0.013	0.019	0.011	0.018	0.016	0.015	0.027	0.021	0.016	0.025	0.012	–
	旅游产业	0.016	0.023	0.016	0.013	0.012	0.016	0.012	0.020	0.015	0.013	0.021	0.012	–
	人才培养	0.010	0.010	0.013	0.010	0.016	0.012	0.011	0.023	0.015	0.015	0.011	0.012	–
	生态建设	0.006	0.004	0.006	0.004	0.006	0.013	0.008	0.011	0.012	0.008	0.009	0.008	↗
	中医药领域的合作交流	0.002	0.003	0.003	0.004	0.004	0.004	0.005	0.005	0.005	0.004	0.005	0.004	↗

（续表 4）

类别	话题	15Q1	15Q2	15Q3	15Q4	16Q1	16Q2	16Q3	16Q4	17Q1	17Q2	17Q3	17Q4	趋势
“一带一路”对我国经济的影响	改革与转型	0.102	0.059	0.084	0.081	0.093	0.059	0.065	0.049	0.062	0.047	0.051	0.048	↘
	区域协调发展	0.055	0.033	0.037	0.047	0.071	0.050	0.046	0.048	0.059	0.026	0.036	0.031	↘
	经济效益	0.034	0.033	0.058	0.030	0.044	0.042	0.033	0.030	0.047	0.029	0.056	0.024	–
	农业供给侧改革	0.015	0.012	0.015	0.017	0.026	0.024	0.018	0.024	0.025	0.017	0.024	0.014	–
政府工作	制度管理	0.053	0.042	0.036	0.033	0.040	0.036	0.039	0.043	0.057	0.046	0.052	0.035	–
	中央精神	0.046	0.014	0.023	0.022	0.037	0.015	0.015	0.029	0.035	0.015	0.032	0.053	–
	政府发展方向	0.038	0.025	0.029	0.027	0.037	0.029	0.025	0.022	0.031	0.022	0.024	0.018	↘

可以看出,不同细分话题的热度变化趋势不尽相同,热度波动类话题占据的比例最大。其中,热度上升类话题包括“会议召开”“进出口”“产权保护”“科研创新”“生态建设”和“中医药领域的合作交流”;热度下降类话题包括“博鳌亚洲论坛”“改革与转型”“区域协调发展”和“政府发展方向”;其余为热度波动类话题。

5 结语

文章基于 LDA 模型对中国政府网“一带一路”相关新闻进行话题抽取及热度演化分析,考察官方媒体对不同话题内容关注度的变化趋势,得出如下结论:

抽取出 30 个“一带一路”相关话题,话题分别属于政策沟通、设施联通、贸易畅通、资金融通、民心相通、“一带一路”对我国经济的影响和政府工作 7 个类别。其中,政策沟通类别的细分话题数最多,包含领导人会谈、博鳌亚洲论坛等 7 个话题,涵盖的内容最为丰富。资金融通类别的细分话题数最少,内容相对单一。从全时间段来看,七大话题类别及 30 个细分话题中,政策沟通、贸易畅通以及“一带一路”对我国经济的影响这三类话题在全时间段上热度较高,占据约 70% 的总热度。从热度演化趋势上看,7 个话题类别的热度整体波动不大,但可以通过每个类别下细分话题的热度演化情况看出官方媒体关注点的变化,例如“改革与转型”等话题属于热度下降类话题,“进出口”等话题属于热度上升类话题,“交通建设”等话题属于热度波动类话题。

文章基于概率主题模型更深入地分析了“一带一路”相关新闻的内容,对当前领域的相关研究进行补充。在未来,可进一步扩展数据源,系统地涵盖各类官方媒体的新闻内容,使结果更为全面,再尝试对新闻话题内容的关联性进行识别,更深入地研究话题内容的演化。

参考文献:

[1] 杜德斌, 马亚华. “一带一路”:中华民族复兴的地缘大战略[J]. 地理研究, 2015, 34(6):1005－1014.
[2] 姚玉娇.《人民日报》“一带一路”专题报道新闻框架研究[D]. 乌鲁木齐: 新疆大学, 2017.
[3] 曾润喜, 魏冯. “一带一路”国家战略的舆论引导评价研究[J]. 情报杂志, 2017, 36(5):90－94.
[4] 汪海藻.《中国日报》与《中国日报·美国版》“一带一路”报道比较研究[D]. 广州:广东外语外贸大学,2017.
[5] 田作宇. 基于语料库的印度英文报纸中“一带一路”相关新闻的态度研究[D]. 北京:北京外国语大学, 2017.
[6] SALTON G, YANG C S. On the specification of term values in automatic indexing[J]. Journal of documentation, 1973, 29(4):351－372.
[7] ALLAN J, PAPKA R, LAVRENKO V. On-line new event detection and tracking[C]// ACM SIGIR Forum. Amherst: University of Massachusetts, 1998:37－45.
[8] 林南. 基于 Web 舆情的话题识别与追踪技术研究[D]. 福州: 福州大学, 2014.
[9] 陈龙. 新闻热点话题发现及演化分析研究与应用[D]. 南京: 南京理工大学, 2017.
[10] LAVRENKO V, ALLAN J, DEGUZMAN E, et al. Relevance models for topic detection and tracking[C]//Proceedings of the second international conference on human language technology research. San Francisco: Morgan Kaufmann Publishers Inc., 2002: 115－121.
[11] ZHAI C, LAFFERTY J. A study of smoothing methods for language models applied to ad hoc information retrieval[C]//Proceedings of the 24th annual international ACM SIGIR conference on research and development in information retrieval. New York: ACM, 2001: 334－342.
[12] HOFMANN T. Probabilistic latent semantic analysis[C]//Proceedings of the fifteenth conference on uncertainty in artificial intelligence. San Francisco: Morgan Kaufmann Publishers Inc., 1999: 289－296.
[13] BLEI D M, NG A Y, JORDAN M I. Latent dirichlet allocation[J]. Journal of machine learning research, 2003, 3: 993－1022.
[14] BLEI D M, LAFFERTY J D. Dynamic topic models[C]//Proceedings of the 23rd international conference on machine learning. New York: ACM, 2006: 113－120.
[15] ALSUMAIT L, BARBARÁ D, DOMENICONI C. On-line LDA: a-

daptive topic models for mining text streams with applications to topic detection and tracking[C]//Proceedings of the 8th IEEE international conference on data mining. Washington: IEEE Computer Society, 2008: 3 – 12.
[16] 楚克明, 李芳. 基于 LDA 模型的新闻话题的演化[J]. 计算机应用与软件, 2011, 28(4):4 – 7.
[17] GRIFFITHS T L, STEYVERS M. Finding scientific topics[J]. Proc. national academy of sciences, 2004, 101(1):5228 – 5235.
[18] 周振宇. 基于 LDA 的微博与传统媒体的话题对比研究[D]. 上海:上海交通大学, 2013.
[19] STEVENS K, KEGELMEYER P, ANDRZEJEWSKI D, et al. Exploring topic coherence over many models and many topics[C]//Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning. Jeju island: Association for Computational Linguistics, 2012: 952 – 961.

作者贡献说明:
秦玥:确定选题,提出论文研究框架,撰写论文;
吴亚平:修改论文,调整框架;
王继民:提出研究思路,修订论文。

An Analysis of News Topics Mining Based on LDA Model: Taking "The Belt and Road" Related News as an Example

Qin Yue[1] Wu Yaping[2] Wang Jimin[1]
[1] Department of Information Management, Peking University, Beijing 100871
[2] Peking University Library, Beijing 100871

Abstract: [**Purpose/significance**] This paper conducted a LDA topic analysis on "the Belt and Road" related news content in official medias and built a basic framework of news topic analysis using LDA model to help the public understand the dynamics and progress of the initiative and its focus in different periods. [**Method/process**] This paper selected "the Belt and Road" related news on the Chinese government Website during 2015 to 2017, and conducted the topic extraction and heat evolution analysis using LDA model. [**Result/conclusion**] A total of 30 topics were extracted and summarized as seven categories called policy coordination, facilities connectivity, unimpeded trade, financial integration, people-to-people bond, economic impact and government work. Among them, the policy coordination category has the highest heat during whole time period. Unimpeded trade category and economic impact category are the second and third highest. The heat of some topics, such as "reform and transformation", decline over time, while others like "import and export" increase. These results reflect the changes in the attention of the official media to different news topics related with "the Belt and Road".

Keywords: "The Belt and Road" LDA model topic extraction heat evolution